

Rapid Methods for Comparing Protein Structures and Scanning Structure Databases

Oliviero Carugo*

Department of General Chemistry, University of Pavia, Italy; and Department of Biomolecular Structural Chemistry, University of Vienna, Austria

Abstract: Databases of three-dimensional macromolecular structures became so large that fast search tools and comparison methods were needed and were actually designed. All of them employ simplified representations of the three-dimensional structure: strings of characters of variable length, which can be handled with procedures that were designed for sequence analysis; fixed dimension arrays that can be processed with standard statistical methods; ensembles of secondary structural elements, which are much less numerous than the atoms/residues of the protein; and continuous representations of the backbone, through stereochemical figures. Some of these computational procedures were developed long ago, when computers were too slow, and others have been designed recently, with the specific aim of handling large amount of information. The present article is focused on the algorithms that allow fast structure comparison, particularly suitable to handle large databases, and should provide a comprehensive picture, useful for the development and the assessment of novel tools.

Keywords: Protein structure, structural alignment, structure comparison, structure database.

INTRODUCTION

Fast procedures for comparing protein three-dimensional (3D) structures are needed. Fig. 1 shows progress in computers and in structural biology during the last three decades. It appears that the computers performance increased as well as data complexity. Poor computers of thirty years ago were used to handle few 3D structures and fast modern computers are used nowadays to handle large amount of protein 3D structures. Structural genomics initiatives [1, 2] will probably provide enormous amount of data or, at least, new techniques to allow highthroughput data production. This requires the development of fast algorithms and protocols to measure similarity between protein 3D structures.

A considerable number of methods are available for comparing protein 3D structures. Some of them (Table 1) were also coded into computer programs that can be used within web-based servers or as stand-alone applications. Why so many methods? The answer is simple: Each biological problem needs its own comparison method. This is not a trivial answer, since it is actually true that different problems need different logical approaches. Some of the reasons why similarity between protein 3D structures must be estimated are summarized, in a schematic way below.

- i) *Molecular Evolution:* Remote homology is detectable more reliably by comparing 3D structures, since 3D features are better conserved than amino acidic sequences [3].

- ii) *Molecular Modeling:* The assessment of 3D structure prediction methods can be performed only by comparing computational models with experimental benchmarks [4].
- iii) *Function Prediction:* The detection of local or global similarity between the 3D structure of protein A, the function of which is unknown, and the 3D structure of protein B, the function of which is known, allows the prediction of the function of protein A [5, 6].
- iv) *Database Scanning:* The Protein Data Bank [7], which is the primary repository of macromolecular 3D structures, and other secondary databases, like for example CATH [8] and SCOP [9], are precious sources of information for both experimentalists and bioinformaticians.

Some of the 3D comparison methods were developed to examine carefully the structural proximity between two or more protein structures. Usually, they result in structural alignments and they are too slow to be used to scan large databases. Often they adopt a two step strategy, where an initial coarse representation of the 3D structure is used to speed up the computations, which are concluded, in the second step, by using a finer structure representation. For example, an initial alignment is obtained by comparing the arrangements of the secondary structural elements (SSEs) and it is subsequently refined by considering the positions of the C α atoms.

Other 3D comparison methods were designed to be faster and, as a consequence, their results are less accurate and detailed. In general, for example, they do not provide structural alignments. They use coarse representations of the protein structure and are suitable to scan large databases. Some of these methods were designed in early times of structural bioinformatics, when computers were much slower than nowadays and algorithms were necessarily simple, from

*Address correspondence to this author at the Institute of Biomolecular Structural Chemistry, University of Vienna, Campus Vienna Biocenter 5, Rennweg 95b, 1030 Wien, Austria; Tel: +43 1 4277 52208; Fax: +43 1 4277 9522; E-mail: oliviero.carugo@univie.ac.at

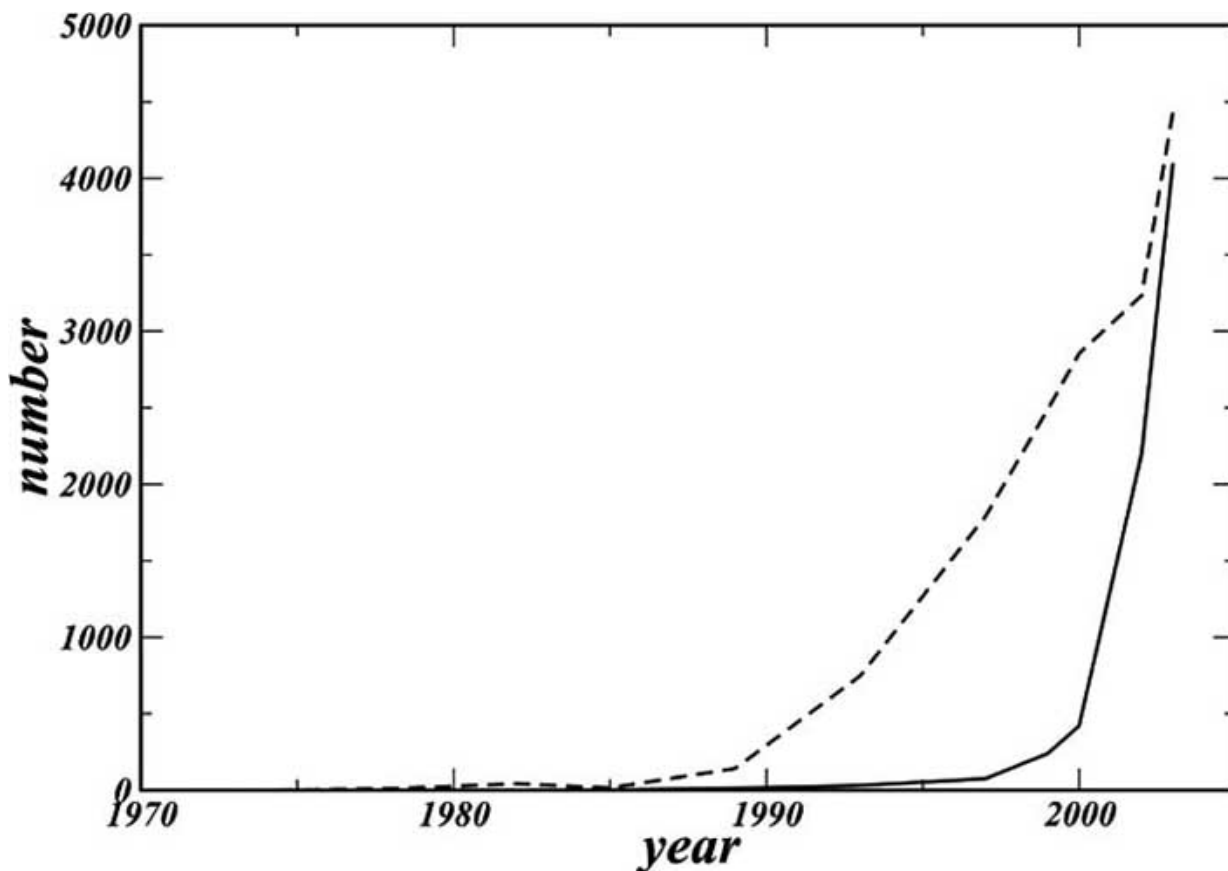


Fig. (1). Increase of computer performance and structural biological information during the last three decades (from 1971 to 2003). The numbers of macromolecular 3D structure available in the PDB are shown by a dashed line. The numbers of transistors per intergrated circuit are shown by a plain line (values $\times 100,000$) (data taken from www.intel.com/research/silicon/mooreslaw.htm). Although this does not describe accurately the relationship between computer power and amount of information, it is a qualitative description of such a relationship. One must in fact consider that: many data in the PDB have been substituted by successive entries and disappeared from the last release; the amount on information within a single PDB file is not constant; the density of transistors does not reflect linearly the cpu power and the computer performance.

a numerical point of view. Others were developed more recently, since even fast modern computers hardly handle the extremely large amount of information that is presently available.

This review is focused on fast comparison methods that can be used to scan large structural databases. It covers both “ancient” techniques, which are nearly forgotten despite their potential usefulness, and recent methods. The various computational procedures are classified according to the 3D structure representation that they use [10]. First the methods based on strings will be summarized, where the 3D structure is represented with one or more series of characters. In such cases, the comparison between two protein structures can be performed through the well assessed techniques used to align protein sequences. Later will be introduced the methods that use arrays of fixed length to represent 3D structures, where the comparison is made by one of the standard techniques for comparing arrays. The comparison methods that represent protein structures with ensembles of SSEs will be presented later. These techniques use several criteria to estimate the similarity between 3D structures and have in common, only the structure representation. A fourth section will be devoted

to the comparison techniques that consider explicitly all the residues along the polypeptide chain. Eventually, the studies that compare various methods will be summarized and commented.

STRING REPRESENTATION

Although relatively uncommon, the representation of a protein 3D structure through a string is appealing, since it allows one to use sequence alignment methods in order to compare two or more 3D structures. In practice, the structure of a protein of n residues (or any other type of structural units) is represented by a string of n characters, each associated with a residue (or structural units). These characters are chosen in an alphabet, each character of which is associated with some structural features. Such an approach is obviously attracting, although it is clearly difficult to design an alphabet that can describe in a exhaustive and unique way, the structural features that are observed in protein 3D structures.

An attempt in this direction, named TOPSCAN, was recently published by Martin [11]. The SSEs are first

Table 1. List of Methods for Comparing Protein Three-Dimensional Structures that are Available as Web-Servers or as Stand Alone Programs

Methods	Address	Reference
Interfaced to structural databases		
3DHit	http://bioinfo.pl/3D-Hit/	[68]
CE	http://cl.sdsc.edu/ce.html	[60]
DALI	http://www.ebi.ac.uk/dali/	[61]
DaliLite	http://www.ebi.ac.uk/~holm/DaliLite/	[69]
DEJAVU	http://xray.bmc.uu.se/usf/	[41, 42]
FATCAT	http://fatcat.burnham.org/	[70]
FoldMiner	http://dlb4.stanford.edu/foldminer/	[71]
MATRAS	http://biunit.aist-nara.ac.jp/matras/	[43, 72]
PRIDE	http://hydra.icgeb.trieste.it/pride/	[23]
SSM	http://protominer.csie.ntu.edu.tw/ProteMiner/	[73]
TOP	http://www.tops.leeds.ac.uk/	[66]
TOPOFIT	http://mozart.bio.neu.edu/topofit/topofit.html	[74]
TOPSCAN	http://www.bioinf.org.uk/topscan	[11]
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml	[47]
Not interfaced to structural databases		
CTSS	http://www.cs.ucsb.edu/~tcan/CTSS	[16]
FlexProt	http://bioinfo3d.cs.tau.ac.il/FlexProt/	[75]
LGA	http://predictioncenter.llnl.gov/local/lga/	[76]
MASS	http://bioinfo3d.cs.tau.ac.il/MASS/	[77]
MaxSub	http://www.cs.bgu.ac.il/~dfischer/MaxSub/MaxSub.html	[78]
MultiProt	http://bioinfo3d.cs.tau.ac.il/MultiProt/	[79, 80]
SARF2	http://123d.ncifcrf.gov/sarf2.html	[81]
SHEBA	http://rex.nci.nih.gov/RESEARCH/basic/lmb/mms/sheba.htm	[82]
STRUCLA	http://asia.genesilico.pl/strucla/	[83]
STRUCTAL	http://molmovdb.mbb.yale.edu/align/	[84]

Each computational method is associated with its URL address and the citation of the literature where it was described.

identified in the protein 3D structure on the basis of the atom co-ordinates, with the STRIDE program [12-14], and vectors are built between their end-points. Depending on the largest component of the vector, which can point in the positive or negative side of the axes x, y, and z, the SSE is associated with one of the twelve characters of Table 2. A protein 3D structure that contains n SSEs is therefore represented by a string of n characters.

Two protein 3D structures can therefore be compared with a standard sequence alignment tool – the algorithm of Needleman and Wunsch [15] in TOPSCAN. This is performed by means of the scoring scheme shown in Table 3 and a gap penalty of 8 and by modifying slightly the scores according to the 3D or sequential proximity of the SSEs and their different solvent accessibility. 24 alignments must be performed for each pair of protein 3D structures in order to permute the direction and orientation of the axes. The best alignment, associated with the best similarity score, is

eventually assumed to measure the degree of similarity between the two protein 3D structures.

Table 2. Set of Characters that can be Associated, in the TOPSCAN Method, with the SSEs as a Function of their Orientation in the 3D Space

Direction	Helix	Strand
+x	B	H
-x	D	J
+y	A	G
-y	C	I
+z	E	K
-z	F	L

Table 3. Scoring Scheme Used by TOPSCAN

Orientation	Same SSE type	Different SSE type
Same	1	3
Different by 1 quadrant	8	1
Different by 2 quadrants	2	0

Another sequence alignment algorithm was used recently by Can and Wang [16], who built a distance matrix and analyzed it with the Smith-Waterman method to find the best alignment [17]. The elements of the distance matrix were computed on the basis of two vectors and a string. The elements of a vector were the curvatures of each residue, the second vector contained the torsions of each residue, and the string elements were the secondary structure states of each residue, determined with the DSSP program [18]. Curvatures and torsions were derived from a spline approximation of the trajectory of the series of ordered C α atoms. The values of the elements of the distance matrix were computed as

$$d_{ij}^{AB} = \sqrt{(\kappa_i^A - \kappa_j^B)^2 + (\tau_i^A - \tau_j^B)^2} + s_{ij}^{AB} \quad (1)$$

where κ_x^y and τ_x^y are the curvature and the torsion of residue X in structure Y and where the variable s_{ij}^{AB} can assume the values of +20 if residue i of protein A and residue j of protein B assume the same secondary structure, or of -20, otherwise.

The two methods summarized above are examples in which the comparison between protein 3D structures is performed through techniques that are widely used in sequence alignments. In both examples, the degree of similarity between two structural units, one from each protein, is defined through scores that resemble those used in sequence alignments, taken, for example, from the PAM and BLOSUM matrices [19-21]. Nevertheless, here the similarity scores between two types of structural units are defined in a rather empirical way and are optimized against a target function based on structural benchmarks, like, for example, independent classifications of structural domains.

ARRAY REPRESENTATIONS

Protein structures can be represented by arrays of real numbers, the dimension of which is independent of the protein dimension. The comparison between two protein 3D structures can therefore be performed by a simple comparison between two arrays of equal length. This feature is particularly interesting, because it is possible to use well assessed mathematical tools, which were developed to measure proximities between objects characterized by the same number of variables. The Euclidean distance between two points in an orthogonal space is a simple example of these tools, though many others were developed for a large variety of purposes [22].

The main problem to solve, if one wants to use fixed length arrays to represent protein structure, is the definition of the arrays, since there is no obvious way to describe an object by means of predefined set of variables. A possible solution of this problem has been proposed by Carugo and

Pongor [23]. Protein 3D structures are represented by the distances between their C α atoms. They are organized accordingly to their sequence distance in such a way that all the inter-atomic distances between the C α (i) and C α ($i+n$) are computed, for $3 \leq n \leq 30$, where n is the number of residues intercalated between the C α (i) and C α ($i+n$) atoms. This results in 28 histograms of distances and the comparison between two protein 3D structures is performed by comparing 28 couples of histograms, in a pairwise way (Fig. 2). Each comparison is performed through a contingency table analysis [24] that allows one to estimate the probability of identity of two distributions. It is therefore possible to estimate 28 values of probability, the average value of which gives the overall probability of identity (PRIDE) between the two protein 3D structures. Such a comparison procedure was found to be able to produce results agreeing with the CATH classification of structural domains [8] and, even more interestingly, was computationally very inexpensive. It was possible in fact to make about 1,000 comparisons per second with a rather primordial SGI R10000 system working at 200 MHz.

SECONDARY STRUCTURAL ELEMENTS

Several methods for comparing protein 3D structures make use of SSEs in order to simplify the description of the structure. While a protein contains several tens or hundreds of residues, it contains, in general, very few tens of SSEs. The consequent drop of the number of variables that must be considered clearly makes the comparisons easier. Often, these methods consist of two steps, the first of which uses SSEs to find an initial alignment that is refined in the second step, where a more detailed and finer representation of the structure, usually based on the position of the C α atoms, is adopted. These methods are actually very different from each other and they are grouped together here only because of their common way to simplify the protein 3D structure. The following paragraphs are therefore focused on the ways used to represent the protein 3D structures by means of their SSEs.

Secondary Structural Assignments

Despite the existence of secondary structures, defined as an ordered and periodical local backbone conformation, was discovered long ago [25], different secondary structural assignments are obtained by using different logical procedures or computer programs. This has already been noted and commented [26, 27] and it is not very surprising since secondary structures were originally defined on the basis of the torsion angles along the backbone, which are actually rather variable. Moreover, a considerable number of procedures, very different from each other, for assigning secondary structure to each amino acidic residue have been designed and used [28].

According to the literature, most of the researchers use DSSP [18] to assign secondary structures. DSSP looks for hydrogen bonds between main-chain atoms (the amido N-H group and the carboxylic oxygen atom) and associates each residue with one of eight types of secondary structure. The most common are the α -helix and the β -strand. Other types of helices (π and 3_{10}) are also recognized and discriminated from the α -helices as well as β -bridges, which are β -strands

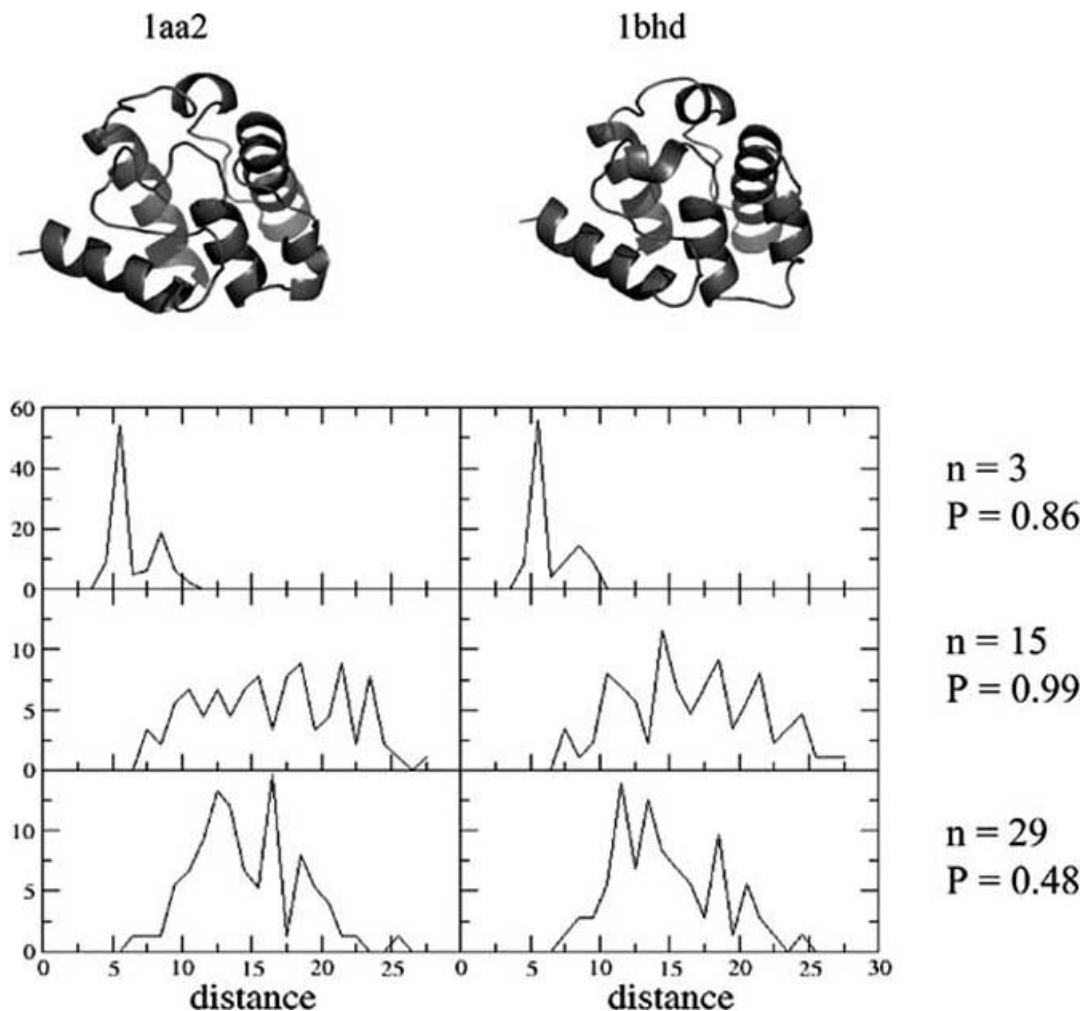


Fig. (2). Example of the procedure for computing the PRIDE scores. The structure of the CH domain of the human β -spectrin (residues 3-107; 1aa2) is compared to the structure of the second CH domain of human utrophin (chain A, residues 151-252; 1bhd). The distributions of the $C\alpha(i)-C\alpha(i+n)$ distances are compared for $3 \leq n \leq 30$. The cases in which $n = 3, 15,$ or 29 are shown. The probability of identity P between two histograms is obtained and all the P values are averaged to give a PRIDE value of 0.86.

formed by only two residue per strand. Also three other types of backbone conformations are detected by DSSP, the turns (single hydrogen bond helices), the bent residues, and the rest.

Also the program STRIDE received considerable attention [12-14]. It is a knowledge based method that uses both hydrogen bonds and backbone torsions to assign secondary structures, by using parameters optimized against a set of assignments made visually by crystallographers and deposited into the PDB [7, 29]. Like DSSP, STRIDE identifies α -, π -, and 3_{10} -helices, β -strands and single residue sheets and, eventually, coil residues that cannot be classified into the other classes.

Other approaches are probably used less commonly, though this remains to be shown. DSSPcont [30] is an extension of DSSP that considers conformational fluctuations of the backbone. P-Curve [31] uses differential geometry to approximate the $C\alpha$ trace of the backbone. DEFINE [32] compares polypeptide moieties to a library of secondary structures. Taylor designed a method based on the

inertial axes of groups of contiguous residues [33]. The method SSA is based on the superposition of polypeptide fragments on sequences of ideal secondary structure [34]. VADAR is a method that considers hydrogen bonding patterns, $C\alpha$ co-ordinate masks, and backbone dihedral angles [35]. Dupuis and co-workers designed a method based on the Voronoi tessellation [36].

All these methods of assigning secondary structure to amino acidic residues in proteins on the basis of the atomic co-ordinates provide somehow contradictory results. DSSP and STRIDE were found to agree in 96% of all residues of a set of 707 non-homologous protein chains and most of the disagreements involved helical assignments [28]. A comparison of DSSP, STRIDE, and DEFINE [27] showed an overall agreement of 71% on a set of 126 proteins. Another comparison between DSSP, DEFINE, and P_Curve [26] showed an overall agreement of only 63% on a data set containing 154 protein chains. DEFINE and P-Curve agreed for 74% of the residues as well as DEFINE and DSSP, while DSSP and P-Curve agreed in 79% of the residues. Because

of these discrepancies, Colloc'h and co-workers proposed a consensus approach based on a majority principle [26]: the secondary structure of a residue is that assigned by the majority of the assignment methods.

It is obvious, on the basis of the considerations reported above, that secondary structure assignments are quite ambiguous and inconsistent. This is particularly true at the borders of the SSEs and it is a serious limitation of the methods that compare protein 3D structures on the basis of the arrangements of the SSEs. In fact, each SSE is usually approximated by a vector that joins its borders.

SSE Approximations

Each SSE is usually described through a vector that goes from its N- to its C-terminus. This can be done in various ways, some of which are simpler, though probably less accurate than others.

For example, Martin simply proposed to join the “end-points of each SSE” [11]. Often, in the absence of pertinent details, this can be interpreted as a vector between the $C\alpha$ atoms of the first and the last residues of the SSE (Fig. 3a). In other cases, it is stated that the beginning and the end of the SSE are the projections on the SSE axis of the $C\alpha$ atoms of the first and last residues (Fig. 3b). On the contrary, Camoglu and co-workers proposed a more sophisticated procedure (Fig. 3c) in which the axis of a SSE span from the center of mass of the first n residues to the center of mass of the last n residues, where $n = 2$ in the case of a β -strand and $n = 4$ in the case of α -helices [37]. All these alternatives in defining the axis of the SSE are clearly prone to produce different results, though this has never been examined in a detailed and systematic way. It can nevertheless be supposed that different approaches result in different SSE

approximations with consequent discrepant proximity scores between protein 3D structures.

Comparison of SSE Arrangements

Once the SSEs have been identified in protein 3D structures, a very wide variety of methods can be used to compare pairs of structures. It is clearly behind the scope of the present review to summarize all methods that were designed and applied to make these comparisons. Only some key aspects are thus commented here.

In most cases, comparisons based on SSEs were only preliminary steps along a multi-step strategy, in which the comparison is performed through structure representations that are increasingly detailed. For example, Yang and Honig [38] obtained an initial and rough 3D alignment between two proteins by comparing the SSE arrangements and refined later this preliminary result by considering explicitly each residue of each protein. Representations based on SSEs are then used to speed up the computations but are not used to measure the real degree of similarity between two protein 3D structures. Similarly, the TOP algorithm [39] is divided into two steps. First, two subsets of SSEs are aligned by monitoring the reciprocal orientation of each pair of SSEs in each structure. Second, the root-mean-square distance of the $C\alpha$ atoms, which were found equivalent in the first step, is minimized. In the method SSM [40], protein 3D structures are represented by graphs, the nodes of which are the secondary structural elements. The comparison between two graphs allows the comparison between two structures and a subsequent minimization of the root-mean-square distance between the residues that are found equivalent after the initial comparison of SSEs allows one to refine the structural alignment. Other similar two-step procedures include DEJAVU [41, 42], Matras [43], and LOCK [44].

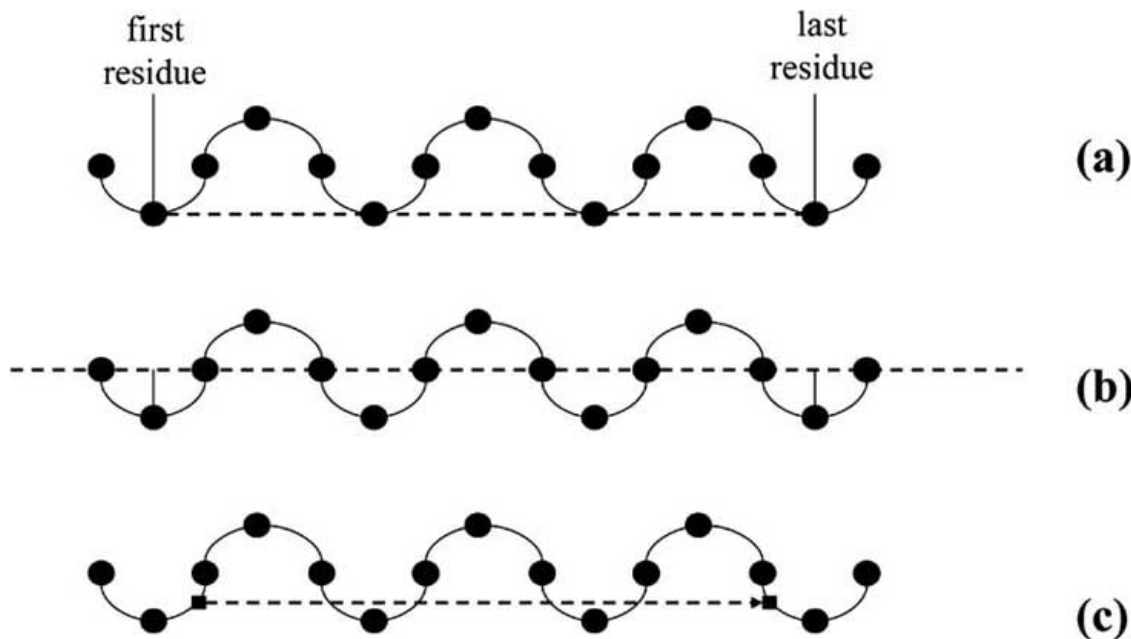


Fig. (3). Three different ways to approximate a SSE with a vector. (a) The vector joins the $C\alpha$ atoms of the first and the last residues of the SSE. (b) The vector joins the projections on the SSE axis of the first and last residue. (c) The vector joins the centers of mass of some of the first and of the last residues of the SSE.

A very wide variety of techniques were used to compare the arrangements of SSEs in pairs of proteins 3D structures and not all of them can be summarized here. Harrison and co-workers [45] represented each structure with a graph, the nodes of which are associated with SSEs and labeled by the type of secondary structure and the edges of which are labeled by the torsion angle and the distance between SSEs. The comparison between two protein 3D structures was then transformed into a comparison between graphs. Two graphs were compared with the Bron and Kerbosch algorithm [46], in order to find common cliques in a fast and effective way. Graph representations, based on SSEs and graph theory algorithms were also used in other tools, like VAST [47] and SSM [40].

Also methods not based on graph theory were developed and tested on practical applications. For example, in the procedure for computing the PSD similarity score, Honig and Yang [38] compared each pair of SSEs of a protein with all pairs of SSEs of the other protein, searching for similar geometrical reciprocal orientations. The 2D array, the elements of which indicate the proximity between a pair of SSEs in one proteins and a pair of SSEs in the other protein, is then processed with a double dynamic programming algorithm [48, 49] that finds the best alignment between pairs of SSEs.

BACKBONE REPRESENTATIONS

Several studies were devoted to methods that describe the trajectory from the N-terminus to the C-terminus along the backbone. The basic idea is to determine a sort of profile of the protein 3D structure in such a way that the comparison between two structures is reduced to the comparison between two profiles. These approaches are very different from those that represent the protein structure with strings, which are compared through techniques usually used to compare protein sequences. In fact, the profile is a sort of vector of real numbers. On the other hand, these approaches differ from those that represent the structure with arrays, since the dimension of the vectors that they use is not fixed. It is thus impossible to use standard matrix algebra and robust statistics for comparing two structures.

The backbone representations are in general rather naïve. These methods define a vector for each protein 3D structure, the elements of which are real numbers and the dimension of which depends on the protein. This makes these methods rather fragile, although they are potentially useful when fast comparison techniques are needed to handle large amount of data. For this reason, they are briefly summarized in the present review.

Pseudo-torsions defined by four consecutive $C\alpha$ atoms have been used by Petsko and co-workers [50] to compare the conformation of a loop in triosephosphate isomerase, which closes over the active site when substrate binds. This procedure, later generalized by Flocco and Mowbray [51], is suited to compare the 3D structures of the same protein in two different conformations, for example with and without a ligand. In practice, it is necessary to compute the torsions defined by the four atoms $C\alpha(i)-C\alpha(i+1)-C\alpha(i+2)-C\alpha(i+3)$ within each structure and compute the differences between these torsions in the two structures. Large absolute values must therefore indicate regions where the two structures

differ. Flocco and Mowbray proposed also a statistical test to find regions where the difference between the two structures is statistically significant, by estimating the expected standard deviation of the torsions on the basis of the average positional error of the $C\alpha$ atoms. A closely related method was designed by Rackovsky and Scheraga, who represented the backbone through a simple curve [52-54]. Each residue is then associated with the curvature and the torsion of the curve and differences of these parameters are used to compare two 3D structures.

A different procedure was developed by Korn and Rose [55], who computed the difference of the ϕ and ψ torsions between two 3D structures of the same protein in two different states. Obviously, large absolute values of these differences are associated with large structural differences between the 3D structures. Such an approach closely resembles that developed by Levine and co-workers [56], who proposed the parameter Δ_{ij} to compare residue i of a protein with residue j of another proteins

$$\Delta_{ij} = |\phi_i - \phi_j| + |\psi_i - \psi_j| \quad (4.2_1)$$

where ϕ_A and ψ_A are the backbone dihedral angles of residue A . Karpen and co-workers developed an improved version of the previous method by comparing all n -length polypeptide moieties of a protein to all n -length moieties of another protein ($3 \leq n \leq 5$) [57].

Another method was developed by Srinivasan and co-workers [58]. It is based on the observation that a protein is constituted by a series of monomers, a moiety of which can be considered to behave as a rigid body. In fact, the atoms N, $C\alpha$, $C\beta$, and C have nearly the same spatial arrangement within each residue. It is thus possible to superpose these atoms belonging to residue i to those belonging to residue $i-1$. This implies the determination of six parameters, three of which are associated with the translation necessary to bring one object over the other and three of which are associated with the rotation necessary to align optimally each pair of equivalent atoms. By plotting the amplitude of the rotation angle versus the residue number, it is possible to obtain a mono-dimensional representation of the protein 3D structure. Conformational differences between two structures can be consequently monitored by computing the differences between the rotation angles.

All the comparison methods summarized above represent the protein structure with vectors, the dimension of which depends on the protein length. It is therefore obvious that the comparison of two 3D structures implies the comparison of these vectors, which may have different dimensions. This is *per se* not impossible, though it is clear that it is hard to handle with gaps and insertions. These comparison procedures are therefore particularly useful when one needs to compare the same protein in two different states, for example with or without the substrate, the inhibitors, the cofactors, etc. It is also possible to use these comparison methods to monitor conformational changes during a molecular dynamics simulation or to evaluate differences between various structural models determined in solution through NMR techniques. On the contrary, these methods can hardly be used in the general case, in which the degree of similarity of two protein 3D structures must be evaluated. It

is nevertheless interesting to keep them in mind, since they are computationally very fast and can therefore be the basis for designing more powerful comparison techniques.

COMPARISONS BETWEEN VARIOUS METHODS

Few extensive and detailed analyses between different methods for comparing protein 3D structures were published. Sierk and Pearson compared [59] seven methods – CE [60], Dali [61], Matras [43], PRIDE [23], SGM [62], Structural [63], and VAST [64]– by using a relatively small number of 86 queries and focused the attention on the quality of the results that each method provides. They found that Dali is the best method to find protein 3D structures, similar to the query, within databases. Kleywegt and co-workers [65] compared 11 publicly available web-servers that allow the user to scan 3D structure databases – CE [60], Dali [61], DEJAVU [42], Lock [44], Matras [43], PRIDE [23], SSM [40], TOP [39], TOPS [66, 67], TOPSCAN [11], and VAST [64] – by using about 70 queries and assessing both the quality of the results that each server provides and the servers speed and overall functionality. CE, Dali, Matras, and VAST showed, on the whole, the best performance, though 100% success rate was achieved by none of the servers.

Not all the computational methods considered in these two comparative analyses can be considered to be fast enough to allow one interactive database scanning [65] and different performances between various tools can be expected because of the different computational approaches. For example, contrary to many other methods, PRIDE can handle only single domain and single chain structures and is inappropriate for other types of proteins. On the other side, PRIDE is suitable to treat structures characterized only by the positions of the C α atoms, while methods relying on SSEs are often inappropriate in these cases.

Although these analyses are interesting for structural biologists, who may use these programs to find protein structures similar to that they are working with, such comparisons tend to be of minor relevance for bioinformaticians interested in computational techniques and in their development. In fact, on one hand, the speed of the servers depends on the computers where the programs are installed and on the number of users that are using the programs at the same time. On the other hand, the results obtained by using different servers and approaches depend on the databases that are scanned and that may be different amongst different web-services. It is thus impossible to exploit the comparisons summarized above in order to compare effectively the performance and functionality of various computational tools designed to browse structural databases. It would be necessary to use bare programs and test them on the same collection of structural data. Such a comparison would nevertheless be extremely useful, despite its inevitable high cost, in order to benchmark the state of the art of fast procedures for comparing protein 3D structures.

REFERENCES

- [1] Gerstein M, Edwards A, Arrowsmith CH, Montelione GT. Structural genomics: Current progress. *Science* **2003**; 299: 1663-3.
- [2] Service RF. Tapping DNA for structures produces a trickle. *Science* **2002**; 298: 948-950.
- [3] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* **1986**; 5: 823-26.
- [4] Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci USA* **2005**; 102: 1029-34.
- [5] Dobson PD, Cai YD, Stapley BJ, Doig AJ. Prediction of protein function in the absence of significant sequence similarity. *Curr Med Chem* **2004**; 11: 2135-42.
- [6] Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **2003**; 36: 307-40.
- [7] Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucl Acids Res* **2000**; 28: 235-42.
- [8] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchical classification of protein domain structures. *Structure* **1997**; 5: 1093-108.
- [9] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of protein database for the investigation of sequences and structures. *J Mol Biol* **1995**; 247: 536-40.
- [10] Eidhammer I, Jonassen I, Taylor WR. Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure, Wiley, Chichester, 2004.
- [11] Martin ACR. The Ups and Downs of protein topology; Rapid comparison of protein structures. *Protein Eng* **2000**; 13: 829-37.
- [12] Heinig M, Frishman D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl Acids Res* **2004**; 32: w500-2.
- [13] Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **1997**; 27: 329-35.
- [14] Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**; 23: 566-79.
- [15] Needleman SB, Wunsch CD. A general method applicable to the search for similarity in the amino acid sequence of two proteins. *J Mol Biol* **1970**; 48: 443-54.
- [16] Can T, Wang Y-F. Protein structure alignment and fast similarity search using local shape signatures. *J Bioinform Comput Biol* **2004**; 2: 215-39.
- [17] Smith R, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* **1981**; 147: 195-97.
- [18] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**; 22: 2577-637.
- [19] Henikoff S, Henikoff JG. Aminoacidic substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **1992**; 89: 10915-19.
- [20] Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. *Proteins* **1993**; 17: 49-61.
- [21] Schwartz RM, Dayhoff MO. Atlas of Protein Sequence and Structure, Nat Biomed Res Found, Washington DC, 1978.
- [22] Theodoridis S, Koutroumbas K. Pattern Recognition, Second edn, Academic Press, San Diego, 2003.
- [23] Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J Mol Biol* **2002**; 315: 887-98.
- [24] Dowdy S, Wearden S. Statistics for Research, Wiley, NY 1991.
- [25] Ramachandran G, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain conformations. *J Mol Biol* **1963**; 7: 95-9.
- [26] Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. Comparison of three different algorithms for the assignment of secondary structure in proteins: the advantage of consensus assignments. *Protein Eng* **1993**; 6: 377-82.
- [27] Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **1999**; 34: 508-19.
- [28] Andersen AFC, Rost B. Secondary structure assignment. *Methods Biochem Anal* **2003**; 44: 341-63.
- [29] Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **1977**; 112: 535-42.
- [30] Andersen C, Palmer A, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. *Structure* **2002**; 10: 175-85.
- [31] Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yieldin complete helicoidal parameters and a unique axis. *Proteins* **1989**; 6: 46-60.
- [32] Richards F, Kundrot C. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins* **1988**; 3: 71-84.

- [33] Taylor WR. Defining linear segments in protein structure. *J Mol Biol* **2001**; 310: 1135-50.
- [34] Craig L, Sanschagrin PC, Rozek A, Lackie S, Kuhn LA, Scott JK. The role of structure in antibody cross-reactivity between peptides and folded proteins. *J Mol Biol* **1998**; 281: 183-201.
- [35] Willard L, Ranjan A, Zhang H, et al. VADAR: A new web server for quantitative evaluation of protein structure quality. *Nucl Acids Res* **2003**; 31: 3316-19.
- [36] Dupuis F, Sadoc JF, Mornon JP. Protein secondary structure assignment through Voronoi tassellation. *Proteins* **2004**; 55: 519-28.
- [37] Camoglu O, Kahveci T, Singh AK. Index-based Similarity Search for Protein Structure Databases. *J Bioinform Comput Biol* **2004**; 2: 99-126.
- [38] Yang AS, Honig B. An integrated approach to the analysis and modelling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol* **2000**; 301: 665-78.
- [39] Lu G. TOP: A new method for protein structure comparisons and similarity searches. *J Appl Cryst* **2000**; 33: 176-83.
- [40] Krissinel E, Heinrick K. Protein structure comparison in 3D based on secondary structure matching (SSM) followed by Ca alignment, scored by a new structural similarity function. Paper presented at the Conference on Molecular Structural Biology, Vienna 2003.
- [41] Madsen D, Kleywegt GJ. Interactive motif and fold recognition in protein structures. *J Appl Cryst* **2002**; 35: 137-39.
- [42] Kleywegt GJ, Jones TA. Detecting folding motifs and similarities in protein structures. *Methods Enzymol* **1997**; 277: 525-45.
- [43] Kawabata T. MATRAS: A program for protein 3D structure comparison. *Nucl Acids Res* **2003**; 31: 3367-69.
- [44] Singh AP, Brutlag DL. Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proc Int Conf Syst Mol Biol* **1997**; 5: 284-93.
- [45] Harrison A, Pearl F, Sillitoe I, et al. Recognizing the fold in a protein structure. *Bioinformatics* **2003**; 19: 1748-59.
- [46] Bron C, Kerbosch J. Algorithm 457 - Finding all cliques of an undirected graph. *Commun Assoc Comput Mach* **1973**; 16: 575-91.
- [47] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* **1996**; 6: 377-85.
- [48] Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. *Proteins* **1992**; 14: 139-67.
- [49] Taylor WR, Orengo CA. Protein structure alignment. *J Mol Biol* **1989**; 208: 1-22.
- [50] Joseph D, Petsko GA, Karplus M. Anatomy of a conformational change: Hinged "lid" motion of the tiophsophate isomerase loop. *Science* **1990**; 249: 1425-28.
- [51] Flocco MM, Mowbray SL. Calpha-based torsion angles: A simple tool to analyze protein conformational changes. *Protein Sci* **1995**; 4: 2118-22.
- [52] Wertz DH, Scheraga HA. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules* **1978**; 11: 9-15.
- [53] Rackovsky S, Scheraga HA. Influence of ordered backbone structure on protein folding. A study of some simple models. *Macromolecules* **1978**; 11: 1-8.
- [54] Rackovsky S, Scheraga HA. Differential geometry and protein folding. *Acc Chem Res* **1984**; 17: 209-13.
- [55] Korn AP, Rose DR. Torsion angle differences as a mean of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. *Protein Eng* **1994**; 7: 961-67.
- [56] Levine M, Stuart D, Williams J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallogr* **1984**; A40: 600-10.
- [57] Karpen ME, Haseth PL, Neet KE. Comparing short protein substructures by a method based on backbone torsion angles. *Proteins* **1989**; 6: 155-67.
- [58] Srinivasan R, Geetha V, Seetharaman S, Mohan S. A unique or essentially unique single parametric characterization of biopolymeric structures. *J Biomol Struct Dyn* **1993**; 11: 583-95.
- [59] Sierk ML, Pearson WR. Sensitivity and selectivity in protein structure comparison. *Protein Sci* **2004**; 13: 773-85.
- [60] Shindyalov IM, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **1998**; 11: 739-47.
- [61] Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* **1993**; 233: 123-38.
- [62] Rogen P, Fain B. Automatic classification of protein structure by using Gauss integrals. *Proc Natl Acad Sci USA* **2003**; 100: 119-24.
- [63] Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci USA* **1998**; 95: 5913-20.
- [64] Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins* **1995**; 23: 356-69.
- [65] Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. *Proteins* **2004**; 54: 260-70.
- [66] Gilbert D, Westhead D, Nagano N, Thornton JM. Motif-based searching in TOPS protein topology database. *Bioinformatics* **1999**; 15: 317-26.
- [67] Gilbert D, Westhead D, Viksna J, Thornton J. A computer system to perform structure comparison using TOPS representations of protein structures. *Comput Chem* **2001**; 26: 23-30.
- [68] Plewczynski D, Pas J, von Grothuss M, Rychlewski L. Comparison of proteins based on segment structural similarity. *Acta Biochim Pol* **2004**; 51: 161-72.
- [69] Holm L, Park J. DaliLite workbench for protein structure comparison. *Bioinformatics* **2000**; 16: 566-67.
- [70] Ye Y, Godzik A. FATCAT: A web server for flexible structure comparison and structure similarity searching. *Nucl Acids Res* **2004**; 32: w 852-8.
- [71] Shapiro J, Brutlag D. FoldMiner: Structural motif discovery using an improved superposition algorithm. *Protein Sci* **2004**; 13: 278-94.
- [72] Kawabata T, Nishikawa K. Protein tertiary structure comparison using the Markov transition model of evolution. *Proteins* **2000**; 41: 108-22.
- [73] Chang DT-H, Chen C-Y, Oyang Y-J, HJuan H-F, Huang H-C. ProteMiner-SSM: A web server for efficient analysis of similar protein tertiary structures. *Nucl Acids Res* **2004**; 32: w76-82.
- [74] Ilyin VA, Abyzov A, Leslin CM. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci* **2004**; 13: 1865-74.
- [75] Shatsky M, Nussinov R, Wolfson HJ. Flexible protein alignment and hinge detection. *Proteins* **2002**; 48: 242-56.
- [76] Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucl Acids Res* **2003**; 31: 3370-74.
- [77] Dror O, Benyamini H, Nussinov R, Wolfson H. MASS: multiple structural alignment by secondary structures. *Bioinformatics* **2003**; 19: i95-104.
- [78] Siew N, Elofsson A, Rychlewski L, Fischer DR. MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **2000**; 16: 776-85.
- [79] Shatsky M, Nussinov R, Wolfson HJ. MultiProt - A multiple protein alignment algorithm in Lecture Notes in Computer Science, Springer Verlag, Heidelberg 2002; 235-250.
- [80] Shatsky M, Nussinov R, Wolfson H. MultiProt - A multiple protein structural alignment algorithm in Workshop on algorithms in bioinformatics. In: Giogo R, Gusfield D Eds, Lecture notes in computer science 2452. Springer Verlag, Rome 2002; 235-250.
- [81] Alexandrov NN. SARFing the PDB. *Protein Eng* **1996**; 9: 727-32.
- [82] Jung J, Lee B. Protein structure alignment using environmental profiles. *Protein Eng* **2000**; 13: 535-43.
- [83] Sasin JM, Kurowski MA, Bujnicki JM. STRUCLA: A www meta-server for protein structure comparison and evolutionary classification. *Bioinformatics* **2003**; 19: i252-4.
- [84] Subbiah S, Laurens DV, Levitt M. Structural similarity of DNA binding domains of bacteriophage and the globin core. *Curr Biol* **1993**; 3: 141-48.